Cavity approximation for graphical models

T. Rizzo,[1] B. Wemmenhove,[2] and H. J. Kappen[2]
[1]*"E Fermi" Center, Via Panisperna 89A, Compendio del Viminale 00184, Rome, Italy*
[2]*Department of Biophysics, Foundation for Neural Networks (SNN), Geert Grooteplein 21, 6525 EZ Nijmegen, Netherlands*
(Received 16 January 2007; published 3 July 2007)

We reformulate the cavity approximation (CA), a class of algorithms recently introduced for improving the Bethe approximation estimates of marginals in graphical models. In our formulation, which allows for the treatment of multivalued variables, a further generalization to factor graphs with arbitrary order of interaction factors is explicitly carried out, and a message passing algorithm that implements the first order correction to the Bethe approximation is described. Furthermore, we investigate an implementation of the CA for pairwise interactions. In all cases considered we could confirm that CA[k] with increasing k provides a sequence of approximations of markedly increasing precision. Furthermore, in some cases we could also confirm the general expectation that the approximation of order k, whose computational complexity is $O(N^{k+1})$ has an error that scales as $1/N^{k+1}$ with the size of the system. We discuss the relation between this approach and some recent developments in the field.

DOI: 10.1103/PhysRevE.76.011102 PACS number(s): 05.50.+q, 05.10.−a, 89.20.Ff, 75.10.Hk

I. INTRODUCTION

The Bethe approximation (BA) is one of the major ingredients leading to the important advances in combinatorial optimization made by the statistical physics community in recent years. The starting point of this line of research can be traced back to (a) the inclusion of the replica-symmetry-breaking scheme in the context of the Bethe approximation [1,2] and (b) the application of the method to single instances [3]. On the other hand, the Bethe approximation has become a key issue in the context of information theory after it was recognized that the well known belief-propagation (BP) algorithm is tightly related to it [4]. This algorithm was introduced in the context of Bayesian networks and has gained interest after the discovery that the fast decoding of Turbo codes and Gallager codes is indeed an instance of BP [5]. Currently the problem of computing the corrections to the BA is attracting increasing attention (see [6–8] for recent literature), not only for the applications mentioned above but also because the BA is the only way of obtaining a mean-field-like solution to many unsolved physical problems, notably Anderson localization.

In this work we reinvestigate the cavity approximation (CA), a tool recently introduced in [6] to study graphical models. The CA is a sequence of approximations defined iteratively such that the BA corresponds to the zeroth order. Its main features are the following for graphical models consisting of N variables: (i) It can be implemented on a given sample (much as the Bethe approximation and at variance with the Replica method), therefore to each approximation corresponds a BP-like algorithm; (ii) the expansion at order k (CA[k]) is correct on graphs with k loops, much as the Bethe approximation is correct on trees; (iii) the computational complexity of the corresponding algorithm grows as N^{k+1}; (iv) when averaged over the samples the CA reproduces the results of the Replica method; indeed, it corresponds to computing the $1/N^k$ corrections within the cavity method. In [6] it was argued that the CA is the natural approximation scheme on locally treelike structures, in the sense that CA[k] yields the $O(1/N^k)$ corrections for models defined on random graphs. In this paper we confirm this expectation by implementing algorithmically the CA; in particular, we apply CA[0] (i.e., BA), CA[1], and CA[2] to instances of graphical models defined on random graphs. We conclude that the CA is an efficient tool to improve (with polynomial complexity) the BA on this class of models that includes notably the error-correcting codes mentioned above. We also formulate the theory in a representation that allows for straightforward generalization to factor graphs with arbitrary order of the interaction factors. Message passing equations for the implementation of such a generalization are given explicitly. We discuss the relationship between this approach and other approaches to go beyond the BA.

II. CAVITY APPROXIMATION: BASIC IDEAS

In [6] the approach was illustrated in the case of binary variables with pairwise interactions. In the following, for the sake of completeness, we present the case of multivalued variables with generic pairwise interactions $H_{ij}(x_i,x_j)$. The same ideas and methods can be applied to models with multiple interactions (factor graphs).

The basic assumption of the BA is that, once a node (say σ_0 in Fig. 1) is removed from the system, the nodes that were connected to it (σ_1, σ_2, and σ_3) become uncorrelated. This is true on a tree but it is not true in general if loops not shown in Fig. 1 are present. From this assumption one can obtain estimates of local averages of the variables. We consider two questions.

(1) How can we estimate the correlation between node σ_2 and σ_3 when node σ_0 is removed from the system?

(2) How can we use these correlations to improve the estimates of the local averages?

In order to answer these questions local cavity distributions are introduced and equations are derived for them. The equations will not be sufficient to compute all the cavity distributions and they will be partially estimated through a Bethe-like approximation. For each node i we define ∂i the

 ©2007 The American Physical Society

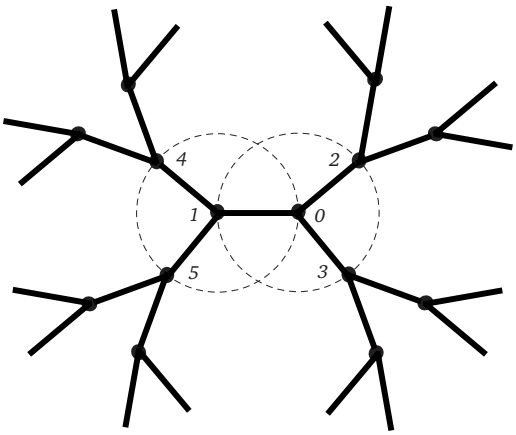

FIG. 1. The marginals of nodes 0 and 1 in the absence of link (01) can be expressed in terms of the joint probabilities of nodes 0,4,5 in the absence of node 1 or of the joint probabilities of nodes 1,2,3 in the absence of node 0. The equality of the results yields the cavity eauations.

neighbors of i and $x_{\partial i} \equiv \{x_j : j \in \partial i\}$. For each node i we consider its cavity distribution, defined as the distribution $P^{(i)}(x_{\partial i})$ of its neighbors in the graph, obtained by removing the variable i from the original graph. Note that the knowledge of $P^{(i)}(x_{\partial i})$ (the Markov blanket of i) is sufficient to determine $P(x_i, x_{\partial i})$ through the formula

$$P(x_i, x_{\partial i}) = c P^{(i)}(x_{\partial i}) \prod_{j \in \partial i} \psi_{ij}(x_i, x_j), \tag{1}$$

where $\psi_{ij}(x_i, x_j) \equiv \exp(-\beta H_{ij}(x_i, x_j))$ and c is a normalization constant. Now we consider the effect of adding to the system without node x_0 all the interactions but ψ_{10}. We can express the marginal of site x_0 in this system in terms of $P^{(0)}(x_{\partial 0})$:

$$P^{(01)}(x_0) = c\Bigg(\sum_{\{x_{\partial 0/1}\}} \prod_{j \in \partial 0/1} P^{(0)}(x_j) \psi_{0j}(x_0, x_j) + \sum_{\{x_{\partial 0/1}\}} \epsilon^{(0)}(x_{\partial 0/1}) \prod_{j \in \partial 0/1} \psi_{0j}(x_0, x_j) \Bigg), \tag{2}$$

where c is a normalization and we have introduced the connected cavity correlation of the set $x_{\partial 0/1}$, $\epsilon^{(0)}(x_{\partial 0/1}) \equiv P^{(0)}(x_{\partial 0/1}) - \prod_{j \in \partial 0/1} P^{(0)}(x_j)$. The same object may be calculated starting from the system without the variable node x_1 and inserting all interactions but $\psi_{10}(x_1, x_0)$:

$$P^{(01)}(x_0) = P^{(1)}(x_0) + \frac{\sum_{\{x_1, x_{\partial 1/0}\}} \epsilon^{(1)}(x_0, x_{\partial 1/0}) \prod_{i \in \partial 1/0} \psi_{1j}(x_1, x_j)}{\sum_{\{x_1, x_{\partial 1/0}\}} P^{(1)}(x_{\partial 1/0}) \prod_{i \in \partial 1/0} \psi_{1j}(x_1, x_j)}, \tag{3}$$

where we have introduced another cavity connected correlation $\epsilon^{(1)}(x_0, x_{\partial 1/0}) \equiv P^{(1)}(x_0, x_{\partial 1/0}) - P^{(1)}(x_0) P^{(1)}(x_{\partial 1/0})$ and the suffix means that quantities are computed in the system without node x_1. Equating the right-hand-side of Eqs. (2) and (3), we obtain an equation that connects the cavity distributions of neighboring nodes:

$$P^{(1)}(x_0) = c\Bigg(\sum_{\{x_{\partial 0/1}\}} \prod_{j \in \partial 0/1} P^{(0)}(x_j) \psi_{0j}(x_0, x_j) + \sum_{\{x_{\partial 0/1}\}} \epsilon^{(0)}(x_{\partial 0/1}) \prod_{j \in \partial 0/1} \psi_{0j}(x_0, x_j) \Bigg) - \frac{\sum_{\{x_1, x_{\partial 1/0}\}} \epsilon^{(1)}(x_0, x_{\partial 1/0}) \prod_{i \in \partial 1/0} \psi_{1j}(x_1, x_j)}{\sum_{\{x_1, x_{\partial 1/0}\}} P^{(1)}(x_{\partial 1/0}) \prod_{i \in \partial 1/0} \psi_{1j}(x_1, x_j)}. \tag{4}$$

We note that this equation is *exact* and is valid also if some of the nodes connected to x_0 coincide with those connected to x_1. We have $2L$ such equations, two for each link [the other equation for link (01) is obtained exchanging indices in Eq. (4) according to $\{0 \leftrightarrow 1,\ 2 \leftrightarrow 4,\ 3 \leftrightarrow 5\}$]. Unfortunately these equations are not sufficient to determine the full set of cavity distributions, which is easily seen noticing that if we knew all the connected cavity correlations $\epsilon^{(i)}(x_j, x_{\partial i/j})$ and $\epsilon^{(j)}(x_{\partial j/i})$ for each link (i,j) then the $2L$ cavity equations should be in principle sufficient to determine the remaining $2L$ unknown cavity distributions $P^{(j)}(x_i)$. The Bethe approximation assumes that the variable nodes on the cavity of node i are uncorrelated in the absence of node i. As a consequence, the corresponding probability distributions are factorized and the connected correlations are zero ($\epsilon^{(i)}(x_j, x_{\partial i/j}) = 0$, $\epsilon^{(j)}(x_{\partial j/i}) = 0$) for each link (i,j); therefore Eq. (4) reduces to the standard belief-propagation equation.

III. ESTIMATING THE CAVITY DISTRIBUTION

In general, if we have an estimate of $P^{(j)}(x_{\partial j})$ for any node j we can compute the various connected correlations in Eq. (4) and solve the cavity equations obtaining an improved estimate of $P^{(i)}(x_j)$. In the following we argue that to estimate the joint probability distribution $P^{(j)}(x_{\partial j})$ it is sufficient to have an algorithm (e.g., BP) that estimates single site marginals $P(x_i)$. Indeed suppose that we have such an algorithm, then in order to get an estimate of $P^{(j)}(x_{\partial j})$ we remove node j from the graph and evaluate $P^{(j)}(x_{j_1})$ through the given algorithm, where $\{x_{j_1}, \ldots, x_{j_k}\} \equiv \partial j$. Then we fix the value of x_{j_1} and compute $P^{(j)}(x_{j_2} | x_{j_1})$ through the same algorithm, and so on. In the end the distribution can be reconstructed from the formula

$$P^{(j)}(x_{\partial j}) = P^{(j)}(x_{j_1}) \prod_{i=2}^{k} P^{(j)}(x_{j_i} | x_{j_1}, \ldots, x_{j_{i-1}}), \tag{5}$$

where k is the number of nodes on the cavity of j. In other words, in order to determine $P^{(j)}(x_{\partial j})$ we have to run the approximate algorithm removing site j and fixing sequentially the values of $x_{\partial j}$. In the following sections we will refer to this procedure to estimate the cavity connected correlations by sequentially fixing the values of the cavity spins as the "clamping" procedure. Note that this is easier to implement than the use of the fluctuation-dissipation-theorem originally proposed in [6] since the latter requires taking derivatives of Eq. (4).

Any algorithm may be used to obtain a first estimate of $P^{(j)}(x_{\partial j})$; in particular, we can use the BA and obtain an im-

proved cavity approximation of order 1 (CA[1]). The procedure can be iterated yielding CA[k] (with CA[0]≡BA) as follows.

(1) Write the exact cavity equations for the system.

(2) For each variable node i:

(a) remove x_i;

(b) express $P^{(i)}(x_{\partial i})$ in terms of conditional probabilities through Eq. (5);

(c) use CA[$k-1$] to compute the conditional probabilities, compute $P^{(i)}(x_{\partial i})$ and then $\epsilon^{(i)}$.

(3) Substitute the estimates of $\epsilon^{(i)}$ into the exact equations and recompute the $2L$ cavity distributions $P_j^{(i)}(x_j)$.

In practice the procedure can be implemented through a message-passing algorithm of which the computational complexity grows with order k as N^{k+1}.

IV. PERTURBATIVE APPROACH FOR PRACTICAL IMPLEMENTATIONS

We note that the use of Eq. (5) requires the application of the algorithm CA[$k-1$] a number of times exponential in the size of the cavities, therefore it may be convenient to use an approximate expression of Eq. (5). In the following we discuss one such approximation. For a given set of nodes $\mathcal{A}$ we define the connected correlation functions as usual, in particular we have: $c(x)=P(x)$, $c(x,y)=P(x,y)-P(x)P(y)$ and so on. The probability distribution of a set of nodes $\mathcal{A}$ can be written as

$$P(x_{\mathcal{A}}) = \sum_{[\mathcal{A}_1,\ldots,\mathcal{A}_n]} c(x_{\mathcal{A}_1})\cdots c(x_{\mathcal{A}_n}), \tag{6}$$

where $[\mathcal{A}_1,\ldots,\mathcal{A}_n]$ runs over the partitions of $\mathcal{A}$. Under some conditions one can assume that $P(x)$ is $O(1)$ while $c(x,y)$ is small, say $O(\epsilon)$ (where ϵ is some small parameter), $c(x,y,z)=O(\epsilon^2)$, etc. For instance, in the representation $P(x_{\mathcal{A}})\propto\exp[\Sigma_i a_i(x_i)+\Sigma_{i<j}a_{ij}(x_i,x_j)+\Sigma_{i<j<k}a_{ijk}(x_i,x_j,x_k)+\cdots]$ this approximation is valid if the interaction terms between k variables are proportional to ϵ^{k-1}. As before the connected correlation functions can be expressed through conditional probabilities, i.e., $c(x,y)=[P(x|y)-P(x)]P(y)$ and can be determined through any algorithm that yields the local distributions $P(x)$. These observations can be used to reduce the number of quantities to be estimated at each cavity, in particular, steps 2(b) and 2(c) can be modified in the following way:

(b) Express $P^{(i)}(x_{\partial i})$ through Eq. (6) assuming that all connected correlation functions of more than $k+1$ nodes vanish.

(c) Use CA[$k-1$] on the corresponding system to determine the connected correlations through conditional probabilities.

In the following we call CA[k] the approximation scheme that includes the previous assumption. It was shown in [6] that CA[k] is exact on graphs with k loops, much as the Bethe approximation is exact on trees. It can be argued that this approximation scheme yields the perturbative expansion in powers of $1/N$ on models defined on random graphs of size N, roughly speaking it means that the CA[k] yields the local marginals with an error $O(1/N^{k+1})$. Indeed in the large N limit random graphs are locally treelike, the loops typically being large. On a locally treelike portion of a random graph the two-point cavity connected correlations are determined by large loops and therefore are small; the three-point cavity correlations depend on the correlations between these large loops and are even smaller, in general we expect that the cavity correlations of k nodes yield an effect $O(1/N^{k-1})$. Therefore, in such a region we expect that CA[k] is really a perturbative expansion. On the other hand, small loops ($l \ll \ln N$) are definitely present in random graphs; see [9] and references therein. The typical graph contains a finite number of small loops and in general graphs with a finite number of small structures of k nearby loops have probability $O(1/N^{k-1})$. Using the exactness of CA[k] on graphs with k loops [6] mentioned above it can be argued that the presence of these small loops does not destroy the perturbative nature of the expansion.

V. GENERALIZATION TO ARBITRARY FACTOR GRAPHS

The above strategy, which up to now has been restricted to two-variable interaction models, may be generalized surprisingly easily for factor graphs with arbitrary number of variables in each factor. We will write down exact equations, as before, for a definition of epsilon functions that corresponds to an expansion around totally factorizing cavity distributions, and later we will neglect higher order terms. The resulting equations explicitly yield a message passing algorithm that takes into account the first order correction (CA[1]) to belief propagation. In our notation, roman indices $(i,j,k,\ldots)$ will denote variables and greek indices $(\alpha,\beta,\gamma,\ldots)$ denote factors. The factor indices are understood to simultaneously represent the subset of roman indices corresponding to the variables in the factor.

A. Exact equations

The exact equations for the marginal of variable x_i in the absence of factor α reads

$$P^{(\alpha)}(x_i) = c\left(\sum_{x_{\partial_i\setminus\alpha}}\prod_{\beta\in\partial_i\setminus\alpha}\left[\psi_\beta(x_\beta)\prod_{j\in\beta\setminus i}P^{(i)}(x_j)\right] + \sum_{x_{\partial_i\setminus\alpha}}\epsilon^{(i)}(x_{\partial_i\setminus\alpha})\prod_{\beta\in\partial_i\setminus\alpha}\psi_\beta(x_\beta)\right), \tag{7}$$

$$P^{(\alpha)}(x_i) = P^{(j)}(x_i) + \frac{\sum_{x_j,x_{\partial_j\setminus\alpha}}\epsilon^{(j)}(x_i,x_{\partial_j\setminus\alpha})\prod_{\beta\in\partial_j\setminus\alpha}\psi_\beta(x_\beta)}{\sum_{x_j,x_{\partial_j\setminus\alpha}}P^{(j)}(x_{\partial_j\setminus\alpha})\prod_{\beta\in\partial_j\setminus\alpha}\psi_\beta(x_\beta)}, \tag{8}$$

where the expansion parameters are given by

$$\epsilon^{(i)}(x_{\partial_i\setminus\alpha}) = P^{(i)}(x_{\partial_i\setminus\alpha}) - \prod_{\beta\in\partial_i\setminus\alpha}\left[\prod_{l\in\beta\setminus i}P^{(i)}(x_l)\right], \tag{9}$$

$$\epsilon^{(j)}(x_i,x_{\partial_j\setminus\alpha}) = P^{(j)}(x_i,x_{\partial_j\setminus\alpha}) - P^{(j)}(x_i)P^{(j)}(x_{\partial_j\setminus\alpha}). \tag{10}$$

B. Truncated expansion

In the following, we will assume that $\forall i \forall \alpha_1, \alpha_2 \in \partial_i$ and we have that $\alpha_1 \cap \alpha_2 = \{i\}$. Up to first order in two-point connected correlations, we may write

$$\epsilon^{(j)}(x_i, x_{\partial_j \setminus \alpha}) \approx \sum_{\beta \in \partial_j \setminus \alpha} \sum_{k \in \beta \setminus i} c^{(j)}(x_i, x_k) \prod_{l \in \beta \setminus i,k} [P^{(j)}(x_l)] \times \prod_{\gamma \in \partial_j \setminus (\alpha,\beta)} \left[\prod_{m \in \gamma \setminus j} P^{(j)}(x_m) \right], \tag{11}$$

$$\begin{aligned}\epsilon^{(i)}(x_{\partial_i \setminus \alpha}) \approx & \sum_{\beta \in \partial_i \setminus \alpha} \sum_{k<l \in \beta \setminus i} c^{(i)}(x_k, x_l) \prod_{m \in \beta \setminus (k,l,i)} [P^{(i)}(x_m)] \\ & \times \prod_{\gamma \in \partial_i \setminus (\alpha,\beta)} \left[\prod_{n \in \gamma \setminus i} P^{(i)}(x_n) \right] \\ & + \sum_{\beta<\gamma \in \partial_i \setminus \alpha} \sum_{k \in \beta} \sum_{l \in \gamma} c^{(i)}(x_k, x_l) \prod_{m \in \beta \setminus (i,k)} [P^{(i)}(x_m)] \\ & \times \prod_{n \in \gamma \setminus (i,l)} [P^{(i)}(x_n)] \prod_{\eta \in \partial_i \setminus (\alpha,\beta,\gamma)} \prod_{r \in \eta \setminus i} [P^{(i)}(x_r)].\end{aligned} \tag{12}$$

Let us introduce some notation:

$$\mu_{\alpha \to i}(x_i) \equiv \sum_{x_{\alpha \setminus i}} \prod_{j \in \alpha \setminus i} [P^{(i)}(x_j)] \psi_\alpha(x_\alpha), \tag{13}$$

$$\lambda_{\alpha \to i}(x_i) \equiv \sum_{x_{\alpha \setminus i}} \sum_{j<k \in \alpha \setminus i} c^{(i)}(x_j, x_k) \prod_{l \in \alpha \setminus (i,j,k)} [P^{(i)}(x_l)] \psi_\alpha(x_\alpha), \tag{14}$$

$$\begin{aligned}\rho_{(\alpha,\beta) \to i}(x_i) \equiv & \sum_{x_{\alpha \setminus i}} \sum_{x_{\beta \setminus i}} \sum_{j \in \alpha \setminus i} \sum_{k \in \beta \setminus i} c^{(i)}(x_j, x_k) \\ & \times \psi_\alpha(x_\alpha) \psi_\beta(x_\beta) \prod_{l \in \alpha \setminus (i,j)} [P^{(i)}(x_l)] \\ & \times \prod_{m \in \beta \setminus (i,k)} [P^{(i)}(x_m)],\end{aligned} \tag{15}$$

$$\nu_{\alpha \to i}(x_i, x_j) \equiv \sum_{x_{\alpha \setminus i}} \sum_{k \in \alpha \setminus i} c^{(i)}(x_j, x_k) \prod_{l \in \alpha \setminus (i,k)} [P^{(i)}(x_l)] \psi_\alpha(x_\alpha). \tag{16}$$

These functions may be interpreted as generalized messages, where the $\mu_{\alpha \to i}(x_i)$ are the familiar ones appearing in belief propagation. Putting things together, we may write up to first order

$$P^{(j)}(x_i) \approx \frac{G^{(\alpha)}(x_i)}{\sum_{x_i} G^{(\alpha)}(x_i)} - \frac{\sum_{\beta \in \partial_j \setminus \alpha} \sum_{x_j} \nu_{\beta \to j}(x_j, x_i) \prod_{\gamma \in \partial_j \setminus (\alpha,\beta)} \mu_{\gamma \to j}(x_j)}{\sum_{x_j} \prod_{\beta \in \partial_j \setminus \alpha} \mu_{\beta \to j}(x_j)} \tag{17}$$

where

$$G^{(\alpha)}(x_i) = \left[\prod_{\beta \in \partial_i \setminus \alpha} \mu_{\beta \to i}(x_i) + \sum_{\beta \in \partial_i \setminus \alpha} \lambda_{\beta \to i}(x_i) \prod_{\gamma \in \partial_i \setminus (\alpha,\beta)} \mu_{\gamma \to i}(x_i) + \sum_{\beta<\gamma \in \partial_i \setminus \alpha} \rho_{\beta,\gamma \to i}(x_i) \prod_{\eta \in \partial_i \setminus (\alpha,\beta,\gamma)} \mu_{\eta \to i}(x_i) \right]. \tag{18}$$

The true marginals

$$P(x_i) = c \sum_{x_{\partial_i}} P^{(i)}(x_{\partial_i}) \prod_{\beta \in \partial_i} \psi_\beta(x_\beta) \tag{19}$$

up to first order read

$$P(x_i) \approx \frac{G(x_i)}{\sum_{x_i} G(x_i)}, \tag{20}$$

$$G(x_i) = \left[\prod_{\beta \in \partial_i} \mu_{\beta \to i}(x_i) + \sum_{\beta \in \partial_i} \lambda_{\beta \to i}(x_i) \prod_{\gamma \in \partial_i \setminus \beta} \mu_{\gamma \to i}(x_i) + \sum_{\alpha<\beta \in \partial_i} \rho_{\alpha,\beta \to i}(x_i) \prod_{\gamma \in \partial_i \setminus \alpha,\beta} \mu_{\gamma \to i}(x_i) \right]. \tag{21}$$

From these equations the pair-interaction case may straightforwardly be recovered: The λ messages do not appear, and the remaining messages are given by

$$\mu_{\alpha \to i}(x_i) = \sum_{x_k} P^{(i)}(x_k) \psi_\alpha(x_i, x_k), \tag{22}$$

$$\rho_{\alpha,\beta \to i}(x_i) = \sum_{x_k, x_l} c^{(i)}(x_k, x_l) \psi_\alpha(x_i, x_k) \psi_\beta(x_i, x_l), \tag{23}$$

$$\nu_{\alpha \to i}(x_i, x_j) = \sum_{x_k} c^{(i)}(x_j, x_k) \psi_\alpha(x_i, x_k). \tag{24}$$

Note that the messages $\lambda_{\alpha \to i}$, $\rho_{(\alpha,\beta) \to i}$ and $\nu_{\alpha \to i}$ should all be "small" compared to $\mu_{\alpha \to i}$ in absolute sense. If they are not, the method is bound to fail in this order of approximation. Indeed these messages are not normalized, contrary to the $P^{(i)}(x_j)$. When the "small" messages are actually neglected in total, it is easily seen that one recovers the belief propagation equations.

C. Complexity issues

In the above form, we can nicely distinguish the dependence of the complexity of the algorithm on factor size versus size of the cavities. The computation of the messages looks exponential in the quadratic factor size, but this cost may be reduced by storing quantities of the form

$$\phi_\alpha(x_i, x_j) = \sum_{x_{\alpha \setminus (i,j)}} \prod_{k \in \alpha \setminus (i,j)} [P^{(i)}(x_k)] \psi_\alpha(x_\alpha). \tag{25}$$

Using these quantities, the computation time scales slightly worse than exponential in the factor size. The dependence on the number of factors in a cavity is found from Eq. (17), from which it is obvious that this dependence is quadratic.

VI. RESULTS: CONFIRMING SCALING OF THE ERROR WITH N

In [6] the perturbative nature of the expansion on random graphs has been confirmed by computing the average of the

energy density in the paramagnetic phase of a spin-glass defined on a random graph. In particular it has been checked that the first order approximation yields the $O(1/N)$ correction to the energy that was computed independently through the replica method.

Currently we report tests of the approximation and the corresponding algorithms on specific instances of random graphs. We have applied the algorithms CA[0] (i.e., BP), CA[1], and CA[2] to systems of binary variables (spin $\sigma_i = \pm 1$, $i \in \{1, \ldots, N\}$) described by the microstate probability distribution

$$P(\sigma) = \frac{e^{\beta \Sigma_{i<j} J_{ij}\sigma_i\sigma_j + \Sigma_i H_i \sigma_i}}{\sum_{\sigma} e^{\beta \Sigma_{i<j} J_{ij}\sigma_i\sigma_j + \Sigma_i H_i \sigma_i}}. \tag{26}$$

The nonzero entries of the matrix J_{ij} form a random graph of fixed connectivity equal to three, and we subsequently investigated ferromagnetic interactions ($J_{ij}=1$ for all nonzero entries) and spin-glass interactions ($J_{ij}=\pm 1$ with equal probability for all nonzero entries).

The use of binary variables allows us to write Eq. (4) in terms of magnetizations and connected correlation functions (see [6]); as explained in Sec. IV, it is assumed that all connected cavity correlations of more than $k+1$ spins vanish when applying the CA[$k-1$] algorithm in the intermediate step of the CA[k] algorithm. In Fig. 2(a) we report the results for a ferromagnet with $H_i=0\,\forall\, i$ at $\beta=0.3$, corresponding to a paramagnetic phase [note that for $N\to\infty$ the critical temperature is given by $\beta_c=\frac{1}{2}\ln(3)$]. We compared the various estimates obtained with CA[0], CA[1], and CA[2] with the exact result obtained through a junction tree algorithm [10]; thus we were forced to consider systems sizes up to $N=120$, although the algorithms we are considering can be applied to much larger systems. For the current ferromagnetic model, a first order loop correction algorithm based on linear response took approximately 110 s to run for a graph of 10^3 nodes, and three hours for 10^4 nodes, on a 2 GHz machine with 1 GB memory (note that the time difference is indeed roughly a factor 10^2 as should be expected). For different sizes of the system we plot the average over 100 random instances of the error of the estimate of the total energy and of the average mean-squared error of the energy of each link. As expected we see that the three algorithms give results of increasing precision; furthermore, we see that the error of the BP (CA[0]) total energy scales with the systems size as $1/N$ while those of CA[1] and CA[2] scale respectively as $1/N^2$ and $1/N^3$.

In Fig. 2(b) we report analogous results for a spin-glass again with $H_i=0$ at $\beta=0.3$ and with random interactions $J_{ij}=\pm 1$. We note that both cases correspond to a paramagnetic region (the critical temperature for the $N\to\infty$ spin glass is given by $\beta_c=(1/2)\ln[(\sqrt{2}+1)/(\sqrt{2}-1)]$), although the algorithm can be applied also in the ferromagnetic region. In the paramagnetic region, however, one may exploit the fact that odd moments of all (marginal) distributions are zero, significantly reducing the complexity of the algorithm. The results for the spin glass model naturally display more fluctuations than the ones for the ferromagnet, since the interaction values are drawn from a distribution, whereas for the ferromagnet they are all equal and thus identical for each of the 100 instances. Since large deviations dominate the average errors for small error values, we plotted error averages in the log domain for the spin glass, i.e., $\exp\langle \ln(\Delta E)\rangle$. Although the correspondence is less convincing than for the ferromagnet, the scaling of errors roughly follows the same exponents. The deviation from this behavior should disappear for larger N. Note that we have used linear response as proposed in [6] to estimate the cavity connected correlations. Details regarding the corresponding update equations for the algorithm are given in the Appendix.

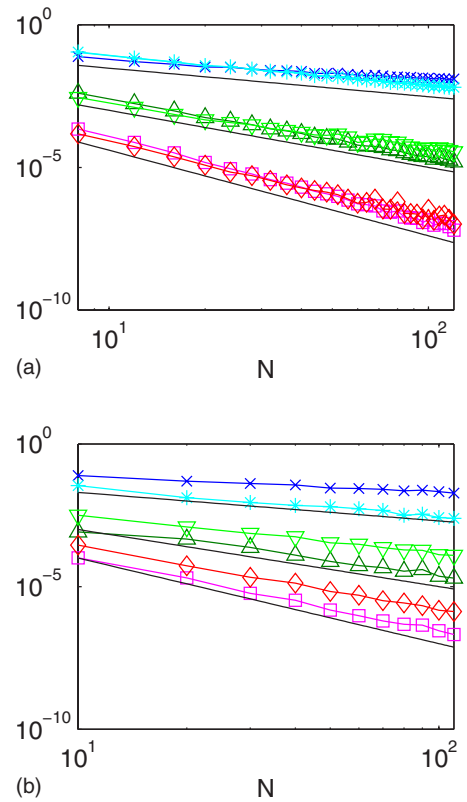

FIG. 2. (Color online) (a) Ferromagnet on random graph with $\beta=0.3$, fixed connectivity 3, and 100 samples. Data points represent averages of the errors of the total energy and mean-squared error per link of the estimates of CA[0] (blue ×, chyan *), CA[1] (light green ▽, dark green △), and CA[2] (red ◇ magenta □), as a function of the sample size N; see text. (b) Same as (a), for a spin-glass model, but average is taken in the log domain (see text). In both figures, black lines with a $1/N$, $1/N^2$, and $1/N^3$ scaling respectively are drawn for comparison.

Figures 3 illustrate the β dependence, where in (a) the total energy is plotted, and in (b) the root mean square error of link energies, both as a function of β, for a ferromagnet with $N=120$. Clearly the CA[1] and CA[2] methods outperform the BA in all but a small region around the "phase transition," where correlation lengths diverge and consequently the connected correlation terms blow up. Naturally perturbative approaches do not result in improved estimates of marginals in this regime, in fact the CA[1] and CA[2] methods cease to converge around $\beta=0.55$. Note that the most difficult part is not estimating the connected correla-

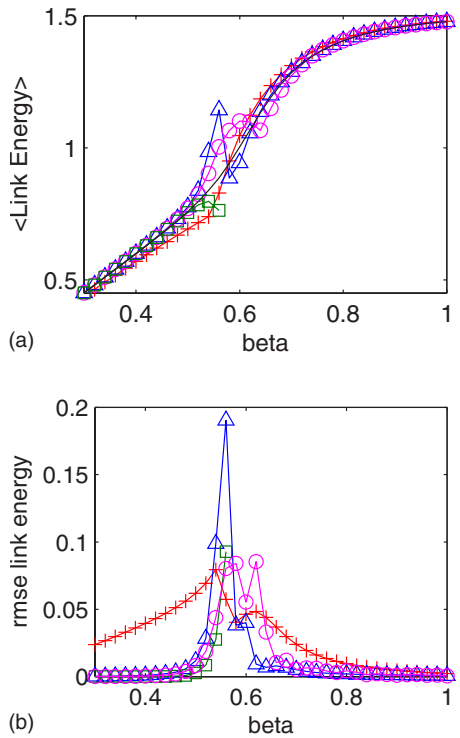

FIG. 3. (Color online) (a) Average energy per link as a function of β for an $N=120$ random ferromagnet as obtained exactly via junction tree (black line), BP or CA[0] (red +), CA[1] with linear response initialization (blue △), CA[1] initialized with the clamping procedure discussed in this paper (magenta ○), and CA[2] (only in the paramagnetic regime, green □). (b) Mean square error of link energy, same color code as (a). Note that in the regime approaching the phase transition the clamping strategy (magenta ○) seems to perform better than the response propagation procedure.

tions, which are based on the BA (and BP converges relatively close to the critical β), but the adapted update equations for $P^{(i)}(x_j)$, Eq. (4) in CA[1] [Eq. (A1) in the Appendix], which do not converge. In general one might be able to optimize an update scheme for these equations, which we have not attempted here. It can be observed in the figure that different ways to estimate the connected correlations may affect the performance near the phase transition, i.e., using linear response results in larger errors than using a clamping procedure as described in Sec. III. Also note that the second order algorithm in the figure is based on linear response.

Apart from problems around the critical value of β, for larger β the inversion symmetry in the model is broken by the approximation algorithms, that consequently disregard the mirror-state free energy valley. In these small-scale models, this does affect the results as can be seen from Fig. 3.

When the symmetry of the model is already broken by a sufficiently large external field, a situation which is common in statistical inference applications, where the external fields may originate from Bayesian priors or represent evidence from measurement data (see, e.g. [11]), this phenomenon does not occur. The CA[1] algorithm consistently improves the marginal estimates over the whole range of β, as illustrated in Fig. 4. When the average external field is relatively

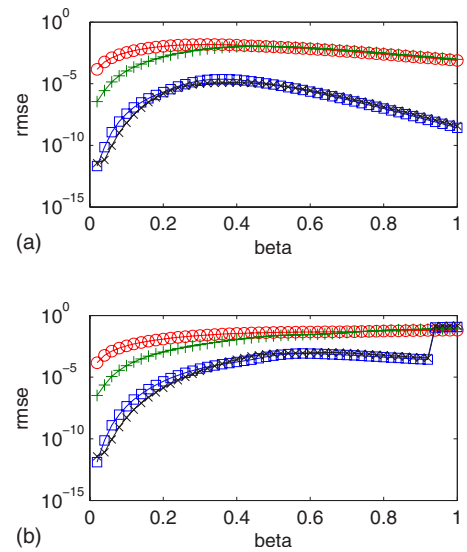

FIG. 4. (Color online) Model of a ferromagnet on a $k=3$, $N=60$ random graph with nonzero magnetizations and random normally distributed external fields of variance 1 and average 1 (a) or 0.2 (b). Root mean square errors of link energies for BP (red ○) are compared to first order corrected BP (blue □). Moreover, root mean square errors of single variable averages for BP (green +) are compared to first order corrected BP ones (black ×). For average external field 0.2 (b) the first order corrected algorithm does not converge for $\beta>0.9$.

small, symmetry breaking might again prevent convergence [Fig. 4(b) for $\beta>0.9$].

In the "magnetized" regime of this model, one may do a similar scaling analysis as displayed in Fig. 2. Results are reported in Fig. 5 for a model with ferromagnetic interactions and different values of the external field average, where we plot the error of the first order CA[1] algorithm as a function of N. Although the results display more fluctuations, the behavior is similar, in the sense that one observes on average a scaling with showing that indeed as long as the parameters correspond to regions not in the vicinity of a phase transition, and the correlation lengths remain typically small compared to the loop length, the approach is promising.

VII. DISCUSSION

The implemented approach is intrinsically perturbative around the BA, in the sense that the BA gives accurate results if the correction terms in Eq. (4) are small and therefore it is natural to guess that CA[1] will produce better results. At the same time if the corrections turn out not to be small, this hints at poor BP estimates, and the whole approach is in trouble [see Fig. 3(b)]. Furthermore we cannot compute CA[1] if BP does not converge. However, we recall that *any* algorithm can be used as the starting point CA[0] of the sequence of approximations. The models for which we produced our proof of principle allow for the implementation of an algorithm in an efficient way, such that averaging over multiple instances of graphs is possible. An algorithmic implementation for more general type of models along the

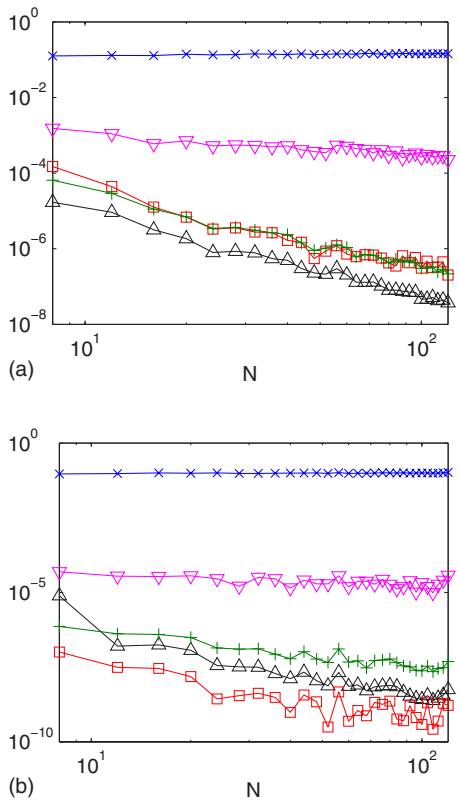

FIG. 5. (Color online) Scaling behavior for a ferromagnet in a broken symmetry magnetized regime caused by normally distributed external fields H_i of variance 1 and average 0.5 (a) versus 1 (b). Blue (×): Root mean square error in BP correlations. Magenta (▽): Root mean square error in BP site magnetizations. Red (□): Root mean square error in CA[1] correlation. Green (+): Root mean square error in CA[1] site magnetizations. Black (△): Error in CA[1] mean energy per link. All data are averages over 100 instances of a random graph of connectivity 3.

lines of Sec. V is more involved, and would require some extra time. We are working on such an algorithmic implementation, but we believe that the results obtained so far are worth publication. Furthermore, the hardest obstacle in this context is the presence of an exponentially large number of large loops, i.e., a nonlocal effect. This crucial problem is already present in the models we considered while considering multivalued variables and interactions modifies the problem at a purely local level and does not change its nature. In conclusion we expect that *whenever BP converges and yields good estimates*, CA[k] *yields a series of approximations of increasing precision. In particular, for graphical models defined on random graphs where small loops are rare*, CA[k] *gives estimates with an error* $O(1/N^{k+1})$ *and with computational complexity* N^{k+1}. Note that this last class of models includes some of the most important present-date error-correcting codes for which the decoding scheme is BP. The reason why BP is so efficient in these cases is precisely that in the corresponding graphical models small loops are rare. Therefore, we expect that by the application of CA[k] the marginals can be computed with any precision in polynomial time. It is important to realize that this does not completely solve the problem of the $1/N$ finite-size effect in error-correcting codes, indeed even if we know the exact marginals, there is still the possibility that some of them are not consistent with the encoded original message.

VIII. RELATION WITH OTHER APPROACHES

The previous comments should help the reader to understand what is the natural context of the present approach and to clarify the relationship with different approaches. A well-known generalization of the BA is Kikuchi's cluster variation method (CVM) [12] which is particularly suitable for finite-dimensional models and in general for models where many small loops are present, indeed this approach amounts to treating loops up to a certain length exactly. On the other hand, on random graphs the corrections to the BA are determined not by small loops (which are rare) but by many large loops. The CVM does not apply to such cases since, in order to include the effect of the large loops, the size of the basic clusters that it treats exactly should be of the order of the total system size, with prohibitive computational complexity. On the other hand, it is natural to expect that CVM performs much better than CA[k] on graphical models defined on structures with many small loops like lattices. Thus the cavity approaches are complementary to CVM, in the sense that both methods have their own well-defined range of applications, although one can imagine applications that could best be studied through a mixture of them.

In a recent publication Chertkov and Chernyak (CC) [7] obtained the free energy of a generic graphical model as an expansion around the BA written in terms of diagrams corresponding to subgraphs with one loop, two loops and so on. In spite of their claim that this represents an improvement with respect to the approach presented here we believe that the two approaches have different motivations and capabilities. The present approach addresses the problem of improving the computation of marginals with polynomial algorithms for models defined on random graphs (with error correcting codes being a notable example of this class of models) and it is as yet not clear if similar results are achievable within the CC approach. Indeed we know that the $1/N$ corrections computed by CA[1] with N^2 complexity are determined by exponentially many large loops (each one yielding a small contribution), therefore it seems likely that in order to obtain results of the quality of CA[1] (i.e., the $1/N$ corrections) one should consider all graphs with one loop in the CC expansion, yielding an exponential number of terms, which is computationally prohibitive unless some resummation scheme is supplemented. Recently [13] an algorithm was tested based on truncation of the series, which may work in cases where one is able to identify the most important loops that contribute to the BP error, when there are not too many.

In a very recent paper [14], a number of different algorithms based on similar ideas as the above have been described, and have been applied to some real-world problems. Given an estimate of the cavity distributions, the update relations in [14] are based on an adjustment of external fields, (keeping higher order interactions in the cavity distribution fixed, whereas, we keep the higher order connected correla-

tions fixed). Although the connection with higher order improvements in their scheme is lost, the algorithms are sometimes easier to implement in the first order case, if the connectivity of the graph is not too large.

Some open problems of the present approach include the computation of the $1/N$ corrections to the free energy (currently we know only how to improve the marginals, therefore we have access only to corrections to local quantities such as the magnetization and the energy) and the extension to the spin-glass phase with the inclusion of replica-symmetry-breaking effects.

ACKNOWLEDGMENTS

We thank Joris Mooij for useful discussions and B.W. acknowledges STW for funding.

APPENDIX: CA[1] AND CA[2] UPDATE EQUATIONS FOR CONNECTIVITY 3

1. CA [1] updates

The update equations for the first moment $M_j^{(i)}$ of a cavity marginal $P^{(i)}(x_j)$ may be written out in terms of connected correlation functions, i.e., $M_k^{(i)}=\Sigma_{x_k}c^{(i)}(x_k)x_k$, $C_{kl}^{(i)}=\Sigma_{x_k,x_l}c^{(i)}(x_k,x_l)x_kx_l$ and $C_{klm}^{(i)}=\Sigma_{x_k,x_l,x_m}c^{(i)}(x_k,x_l,x_m)x_kx_lx_m$ (see [6]):

$$M_j^{(i)}=\frac{t(H_j)K_{ij}+\sum_{l\in\partial j\backslash i}t_{jl}M_l^{(j)}}{K_{ij}+t(H_j)\sum_{l\in\partial j\backslash i}t_{jl}M_l^{(j)}}-\frac{L_{ij}+t(H_i)\sum_{k\in\partial i\backslash j}t_{ik}C_{jk}^{(i)}}{K_{ji}+t(H_i)\sum_{k\in\partial i\backslash j}t_{ik}M_k^{(i)}}, \tag{A1}$$

where

$$K_{ij}=1+\sum_{k<l\in\partial j\backslash i}t_{jk}t_{jl}[M_k^{(j)}M_l^{(j)}+C_{kl}^{(j)}], \tag{A2}$$

$$L_{ij}=\sum_{k<l\in\partial i\backslash j}t_{ik}t_{il}[C_{jl}^{(i)}M_k^{(i)}+C_{jk}^{(i)}M_l^{(i)}+C_{jkl}^{(i)}]. \tag{A3}$$

The solution of these equations leads to the moment of the true marginal $P(x_i)$ via

$$M_i=\frac{T_{\text{odd}}^{(i)}+t(H_i)T_{\text{even}}^{(i)}}{t(H_i)T_{\text{odd}}^{(i)}+T_{\text{even}}^{(i)}},$$

$$T_{\text{odd}}^{(i)}=\sum_{l\in\partial_i}t_{il}M_l^{(i)}+\sum_{(l,k,m)\in\partial i}t_{il}t_{ik}t_{im}M_l^{(i)}C_{km}^{(i)}+\sum_{l<k<m\in\partial_i}t_{il}t_{ik}t_{im}[M_l^{(i)}M_k^{(i)}M_m^{(i)}+C_{lkm}^{(i)}],$$

$$T_{\text{even}}^{(i)}=\left[1+\sum_{k<l\in\partial_i}t_{il}t_{ik}[M_l^{(i)}M_k^{(i)}+C_{lk}^{(i)}]\right]. \tag{A4}$$

Correspondingly, the nearest neighbor correlations read

$$\sum_{x_i,x_j}P(x_i,x_j)x_ix_j=\frac{F_{ij}}{T_{\text{even}}^{(i)}+t(H_i)T_{\text{odd}}^{(i)}}, \tag{A5}$$

where

$$F_{ij}=t(H_i)[M_j^{(i)}+L_{ij}]+\sum_{l\in\partial_i\backslash j}t_{il}[C_{jl}^{(i)}+M_j^{(i)}M_l^{(i)}]+t_{ij}\left[K_{ji}+t(H_i)\sum_{l\in\partial i\backslash j}t_{il}M_l^{(i)}\right]+t(H_i)\sum_{l<k\in\partial i\backslash j}t_{ik}t_{il}M_j^{(i)}[M_l^{(i)}M_k^{(i)}+C_{lk}^{(i)}]. \tag{A6}$$

In the CA[1] approximation, the two-point connected correlations are estimated by some algorithm, possibly CA [0] (another option is to use response propagation, see [6]), and the three-point connected correlations are neglected.

2. CA[2] updates

The CA[2] algorithm in turn uses improved estimates of the two-point connected correlations of which the accuracy corresponds to CA[1], together with CA[0] (or response propagation) three-point estimates. We used response propagation to compute the CA[1] accurate two-point cavity connected correlations. This implies we exploit

$$C_{ik}^{(j)}=\beta^{-1}\frac{\partial M_i^{(j)}}{\partial H_k}, \tag{A7}$$

but $M_i^{(j)}$ is computed with CA[1] accuracy, i.e., from Eq. (A4) on the graph from which variable j has been removed. This may be achieved by simply taking the derivative of the right-hand side of Eq. (A4). In this expression, we encounter $\partial M_k^{(i)}/\partial H_l$ and $\partial C_{kl}^{(i)}/\partial H_m$. The first may be found from the iterative equation resulting from taking the derivative of Eq. (A1), the second may be estimated with a CA[0] or response propagation algorithm, since it is of the order of $C_{klm}^{(i)}$.

In the paramagnetic phase of pair interaction networks without external field, simplifications occur since we may exploit the fact that odd moments of distributions are zero. Consequently, all terms M_i, $M_i^{(j)}$, $C_{ikl}^{(j)}$, and $\partial C_{jk}^{(i)}/\partial H_l$ vanish and the recursive update relations for the derivatives of Eq. (A1) reduce to

$$\frac{\partial M_l^{(j)}}{\partial H_n}=\left[\frac{\sum_{k\in\partial l\backslash j}t_{lk}\frac{\partial M_k^{(l)}}{\partial H_n}}{1+\sum_{k<r\in\partial l\backslash j}t_{lk}t_{lr}C_{kr}^{(l)}}+\delta_{ln}\beta\right]-\frac{\sum_{k<r\in\partial j\backslash l}t_{jk}t_{jr}\left[\frac{\partial M_r^{(j)}}{\partial H_n}C_{lk}^{(j)}+\frac{\partial M_k^{(j)}}{\partial H_n}C_{lr}^{(j)}\right]}{1+\sum_{k<r\in\partial j\backslash l}t_{jk}t_{jr}C_{kr}^{(j)}}+\delta_{jn}\frac{\beta\sum_{k\in\partial j\backslash l}t_{jk}C_{lk}^{(j)}}{1+\sum_{k<r\in\partial j\backslash l}t_{jk}t_{jr}C_{kr}^{(j)}}. \tag{A8}$$

The solution of these equations is to be substituted in

$$\frac{\partial M_j}{\partial H_n} = \frac{\sum_{l \in \partial j} t_{lj} \frac{\partial M_l^{(j)}}{\partial H_n}}{1 + \sum_{k<l \in \partial j} t_{jk} t_{jl} C_{lk}^{(j)}} \tag{A9}$$

on the graph without variable i, yielding a CA[1] computation of $C_{jn}^{(i)}$. This results in improved CA[2] estimates of correlations via Eq. (A6) which simplifies greatly due to the vanishing of odd moments, i.e.,

$$\sum_{x_i,x_j} P(x_i,x_j) x_i x_j = t_{ij} + \frac{\sum_{k \in \partial_i \backslash j} t_{ik} C_{jk}^{(i)}}{1 + \sum_{k<l \in \partial_i} t_{ik} t_{il} C_{kl}^{(i)}}. \tag{A10}$$

[1] M. Mézard and G. Parisi, Eur. Phys. J. B **20**, 217 (2001).
[2] M. Mézard and G. Parisi, J. Stat. Phys. **1**, 111 (2003).
[3] M. Mézard, G. Parisi, and R. Zecchina, Science **297**, 812 (2002).
[4] J. S. Yedidia, W. T. Freeman, and Y. Weiss, *Understanding Belief Propagation and its Generalizations* in *Exploring Artificial Intelligence in the New Millennium* (Morgan Kaufmann, San Francisco, CA, 2003); *Advances in Neural Information Processing Systems 13* (MIT Press, Cambridge, MA, 2000), pp. 689–695.
[5] R. McEliece, D. MacKay, and J. Cheng, IEEE J. Sel. Areas Commun. **16**, 140 (1998).
[6] A. Montanari and T. Rizzo, J. Stat. Mech.: Theory Exp. 2005, P10011 (2005).
[7] M. Chertkov and V. Y. Chernyak, Phys. Rev. E **73**, 065102 (2006).
[8] G. Parisi and F. Slanina, J. Stat. Mech.: Theory Exp. 2006, L02003 (2006).
[9] E. Marinari and R. Monasson, J. Stat. Mech.: Theory Exp. 2004, P09004 (2006).
[10] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, San Francisco, CA, 1988).
[11] D. J. Mackay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, Cambridge, UK, 2003).
[12] A. Pelizzola, J. Phys. A **38**, R309 (2005).
[13] V. Gomez, J. Mooij, and B. Kappen, e-print arXiv:cs.AI/0612109.
[14] J. Mooij and B. Kappen, e-print arXiv:cs.AI/0612030.